



非構造化データ管理のためのベストプラクティス



TABLE OF CONTENTS

- 2 エグゼクティブサマリー
- 3 データ管理の進化
- 5 非構造化データの管理および分析における課題
- 7 効果的なソリューションとは
- 8 非構造化データ管理のための統合されたグローバルシステム
- 9 Snowflakeについて

エグゼクティブサマリー

非構造化データは膨大な量にのぼり、しかも増え続けています。Computer Weekly誌によると、ビジネスに関連する情報の5分の4(ほとんどがEメール、レポート、記事、カスタマーレビュー、顧客ノート、ソーシャルメディアへの投稿などのテキストデータであり、そのほか音声、動画、リモートシステムの監視データなどを含む)が非構造化データであるとのことです。¹

非構造化データは、従来のデータ管理ツールを使ってデータから価値を引き出そうとする企業にとっては多くの課題を伴います。まず検索、分析、クエリを簡単に実行することができず、特に転送中はその傾向が顕著になります。またデータ自体が複雑であるため分析的なインサイトを得るための処理に問題が生じます。さらに可視性とコントロール性の欠如が、ガバナンスやデータセキュリティに関する様々な問題を引き起こします。

非構造化データを(構造化ファイルや半構造化ファイルと併せて)効果的に取り込むことができる最新のデータ管理プラットフォームは、より完成されたデータアナリティクスを可能にしたり、また意思決定のための効果的なインサイトが得られるなど重要なメリットを提供します。効果的なソリューションは、データサイロの排除、迅速かつフレキシブルなデータ処理、シンプルかつセキュアなアクセス性という3つの中核的機能を備えている必要があります。



データ管理の進化

企業や政府などがコンピュータに投資する理由は、そのデータの分析力にあります。戦術的かつ戦略的に優位に立つためのインサイトを引き出すことが常に目標とされています。初期のコンピュータは専ら、当時入手可能だった少量の生データを基に長く難しい数学の問題を解くために使われていました。

しかし現在では、さまざまなソースから大量のデータが入手でき、その形式も構造化、半構造化、非構造化などさまざまです。従来のデータ管理技術は複数のデータ形式全てに対応できないため、企業はすべてのデータから最大の価値を得るための新たな方法を模索し始めています。

どんな形式のデータも等しく重要であり、完全な分析結果を得るにはすべての形式のデータをカバーしなければなりません。

構造化データ

従来のデータ管理システムはいずれも数十年前に設計されたものです。その当時はデータが予測しやすい構造化された形式で送られてきており、データソースが限定的で変更も頻繁には行われていなかったことから、固定スキーマのリレーショナルデータが主流でした。テーブルベースのデータウェアハウスは、このようなデータを保存し管理するために高度にコントロール

された環境でした。この時点では、データがしっかり整理されており分析アルゴリズムによって読み取るのも簡単だったことから、ほとんどのデータアナリティクスは構造化データに限定されていました。

半構造化データ

その後、データ保存コストの急速な低下と分散型システムの発達により、機械生成によるデータが爆発的に増加しました。JSONやAvroなどの半構造化データがデータの送信や保存における標準的な形式となりました。これらデータは、その生成においてもプログラムベースの後処理においても機械処理がしやすいよう意図されています。

一般的に半構造化データは、人の手で開発されたテーブルベースのデータ管理システムのようなテーブル形式の構造にはなっていませんが、タグやその他のマーカーを挿入することでセマンティックな要素の分離や階層化を可能にしています。² この10年でデータレイクが登場したことにより半構造化データ



をより簡単に管理できるようになりました。最近では、テーブルベースの管理システムとファイルベースの管理システムを併用している組織もあります。

非構造化データ

データレイクを利用するとより多くの種類のデータの管理と分析が可能となりますが、現在企業が収集しておりその量が急増している非構造化データにはデータレイク型のアーキテクチャではうまく対処できません。分析を要する非構造化データは急激に増加しています。Analytics Insight社が報告したIDCの予測によると、2025年には世界のデータの80%が非構造化データになると言われていますが、今の段階で分析され利用されているのはそのうちのわずか0.5%とのことです。³

非構造化データは、人の手によって自然に生み出されるものです。機械が周りの世界とやり取りすることで膨大な量の半構造化データが生まれる一方で、人が組織とやり取りすることで膨大な量の非構造化データが生まれます。Wikipediaによると、非構造化データは「あらかじめ定義された方法で整理され

ていないデータ」として定義されています。決められた方法で整理されていないために不規則性や曖昧性が生じ、従来のアプローチでは管理やセキュリティの確保、ガバナンス、処理が難しくなります。⁴

非構造化データには、画像、動画、音声、PDFドキュメントなど複雑なデータを格納したデジタルファイルが含まれます。DICOM (医療画像)、.vcf (ゲノミクス)、.kdf (半導体)、.hdf5 (航空宇宙) など、様々な業界固有のファイル形式もこれに含まれます。

非構造化データは一般に、顧客分析やマーケティングインテリジェンスアプリケーションに利用可能であるが未だ実際に活用されていないリソースとして認識されています。非構造化データからの価値の抽出には大きな可能性がありますが、その複雑さと生成される情報量の多さから、まず非構造化データの管理方法を革命的に進化させる必要があるでしょう。企業は、自らが保有する非構造化ファイルにスムーズにアクセスし、処理し、管理できる方法を必要としています。

非構造化データはどのように役立つか

非構造化データを分析や意思決定に効果的に取り入れることができれば、組織に新たな視点をもたらし、新たなビジネスチャンスを生み出せます。非構造化データで実現できることの例をいくつかご紹介します。

- ソーシャルメディア上で特定の製品について話している顧客の地域等の属性を特定することでソーシャルメディア上の顧客行動を分析し、ターゲットを絞ったマーケティングキャンペーンを展開する。
- 画像ファイルに機械学習 (ML) を自動的に適用してパターン認識を実行することにより、自動車保険の申請処理を迅速化する。
- コールセンターの音声記録を分析し、センチメント分析などのマーケティングインサイトを引き出す。
- 医師の手書きメモをスキャンして臨床試験の候補としての適性を示す用語を抽出し、その情報を構造化データと結合することで臨床試験の候補者をより早く特定して登録できるようにする。

非構造化データの管理および分析における課題

従来のデータ管理システム（データウェアハウスやデータレイク）では、現在のデータ量、速度、多様な形式に対するすべてのワークロード要求に対応しきれません。結果としてこれらのシステムには、さまざまなタイプのデータ（構造化、半構造化、非構造化）に対応する目的でそれぞれ異なるタイプのツールを追加しなければなりません。

DCIGは、「現在、多くの組織が複数のペタバイトのデータを管理する必要に迫られています。ペタバイト規模になると、保存、保護、バックアップ、リカバリーのすべてがレガシーソリューションの処理では問題となります」と指摘しています。⁵

パブリッククラウドプロバイダーが提供するBlobストレージサービス（Amazon S3やAzure Blob Containersなど）が、非構造化データファイルのデフォルトのストレージとなっています。しかし、分析のユースケースに関しては、多くの制約があります。例えば、Blobストレージ内のファイルのリストアップは難しく、実行できるのはプレフィックススペースの検索に限られています。データ保存の指針となる正式なテーブルベースやファイルベースの整理システムがなければ、これらのサービスによる非構造化データへの一貫したアクセス、管理、コントロール、検索、およびセキュリティの確保は非常に困難になります。

非構造化データの複雑性

非構造化データは、本質的に複雑で、分析しにくいものです。保存されている非構造化データはさまざまに異なるファイル形式で構成されているため、集められた情報を意味のある形で把握することは容易ではありません。

音声や動画のメディアファイルを含む非構造化データの場合、他のデータ形式やデータセットとの結合も困難です。これらの問題点により、サイロ化が生じたりデータが活用しきれないといった状況に陥ります。データがサイロ化していると、可視性が低くなることでクエリの効果が限定的となり、一部のデータにはまったくアクセスできなくなります。

データ処理の問題点

異なるバラバラなデータ管理ツールやシステムに依存していると、データパイプラインが複雑で入り組んだものとなり、分析の性能が低下します。PDFファイルからテキストを抽出したり、画像認識ソフトを使って非構造化データを構造化データに変換する作業は煩雑で計算負荷が高く、処理時間が長くなります。非構造化データの管理をレガシーソリューションに依存した状態では、データパイプラインの不具合や、データ移動ではデータを頻繁に別の場所にコピーするためエラーが発生しやすいといった処理上の問題が発生します。こうした障壁はデジタルイノベーションを阻害し、意図したデータ運用効果を達成できないため、組織の目標を実現できなくなります。



ガバナンスとセキュリティの不確定要素

大量の複雑な非構造化データを従来のデータシステムの硬直したアーキテクチャで処理しようとする、データアクセスの管理が非常に困難になります。これは、データの種類やユーザーの役割に応じてアクセスしなければならない場合に特に顕著となります（「ゼロトラスト」のセキュリティ管理要件）。

Security Weekly誌によると、政府のサイバーセキュリティの専門家らは、国のサイバーセキュリティ体制を強化するための最も直接的かつ実用的な方法として、クラウドへの移行とゼロトラストアーキテクチャの導入の2つを明確に打ち出しています。⁶

EUの一般データ保護規則（GDPR）などのデータプライバシー法では、構造化データと非構造化データを区別していません。個人情報を含むデータは、その形式に関係なく常に組織のコントロール下に置き保護しなければなりません。CPO Magazineによると、GDPR違反の罰金は2020年に39%急増しており、EU加盟国の2021年1月時点での罰金総額は約3億3,240万米ドルとなるとのことです。⁷

非構造化データをめぐるガバナンスとセキュリティの具体的な課題は以下の通りです。

Gartner社は、「2023年には世界人口の65%の個人データがプライバシー規制の対象となる（2020年のこの割合は10%）」と予測しています。⁸

- **既存の許可の移行の難しさ。**非構造化データは他のプラットフォームから取得されることが多く、保存されるファイルにはすでにそれらのシステムに関連した複雑な許可が設定されています。それらの権限を把握するだけでも複雑な作業であり、そのうえそれらを新しいプラットフォームにマッピングすることは非常に困難です。
- **データシェアリング。**Verizon社によると、昨年データ漏洩の61%には認証情報が含まれており、特に25%では盗まれた認証情報が使用されていたとのことです。⁹ 企業は、ユーザーに認証情報を与えずにどうやってデータへのアクセスを可能にできるでしょうか？
- **データ移動に伴うリスク。**複数の場所にデータのコピーが存在するデータのサイロ化により、多くの不要なリスクが発生します。
- **忘れられる権利。**アクセスできないデータや異なる管理アーキテクチャからコピーされたデータがあると、国や地域のデータプライバシー法順守のための完全な消去が困難になる場合があります。その結果、規制当局からの罰金や訴訟費用のリスクが発生します。



効果的なソリューションとは

非構造化データの保存と管理は、データアーキテクチャ管理者にとって最も重要な業務の1つです。非構造化データを管理するための効果的なソリューションには、増え続けるデータの保存ならびにそれらへのアクセス、処理、管理、セキュリティ、共有を可能にする機能が組み込まれている必要があります。具体的には、十分な性能、同時実行性、スケーラビリティの確保とともに、利用しているレガシーアプローチの重大な欠点を解消する能力が求められます。

データサイロの排除

現代のデータ管理は、一元化されたクラウドベースのプラットフォームをベースとしながらあらゆるデータ形式（構造化、半構造化、非構造化）に対応し、ファイルの保存、アクセス、処理、共有、分析をスムーズに行えるものでなければなりません。データエンジニアは、どのような種類のクラウドからでもファイルの保存や取得を行える必要があります。それにより、複数のクラウドや異なる地域間でデータへのアクセスが可能となり、統一されたポリシーを適用することができます。

シンプルに合理化されたアーキテクチャを採用することで、メンテナンスや管理のコストが軽減されます。また、非構造化データファイルを必要に応じて内部ステージまたは外部ステージにフレキシブルに保存できるようになります。

迅速かつフレキシブルな処理

現代のデータ管理ソリューションには、複雑な分析、データサイエンス、およびインタラクティブなアプリケーションを使用して完全なインサイトを引き出すために非構造化データの変換、準備、強化に対応できる十分な処理能力が求められます。また手動での微調整を必要とせず、ワークロードの競合も生じさせ

ず、スピーディーで信頼性の高い性能を提供する必要があります。さらにはユーザーやジョブの件数やデータ量にかかわらずコスト効率の高いスケーラビリティを備えることにより、フレキシブルな同時実行性を提供しなければなりません。

コンピュータ性能に加えて、データサイエンティストが最適な生産性で非構造化データを処理するためにはそれぞれが使い慣れたツールを自由に利用できる必要があります。また、継続的なデータパイプラインを確保するために、他の人も簡単に利用できるような透明性を出力データに持たせる必要があります。

簡単にセキュアなアクセス

最後に、データ管理ソリューションはユーザーが非構造化データを簡単に検索し共有できるものでなければなりません。また、それぞれのステージでファイルをすぐ見つけ出せるようファイルカタログを内蔵している必要があります。さらに、範囲を限定したアクセスに対応している必要もあります。これは、物理的なコピーを作成したり、物理ファイルにアクセスするための認証情報を共有したりすることなく、カタログ上に安全なビューを作成し、そのセキュアなビューを他のアカウントと共有できるようにするためです。

企業は、保有するデータに沿ったフレキシブルなポリシーを全てのユーザーやワークロードに対して適用するための包括的なガバナンスを必要としています。ゼロトラスト要件を維持するため、データ管理ソリューションは、それぞれの定義されたユーザーロールに応じて機密データへのアクセスを適切にコントロールしなければなりません。これを実現するために、非構造化ファイルのガバナンスには、あらゆる種類のクラウドに対応するロールベースのアクセスコントロール（RBAC）コマンド（シンプルなGRANTやREVOKEステートメントなど）の利用が求められます。これにより、クラウドプロバイダーのIAM（Identity and Access Management）システムにおけるセキュリティやガバナンスのポリシーの潜在的な複雑性を回避できます。



非構造化データの管理のための統合された グローバルシステム

これから目指すべきデータ管理の新たなステージは、あらゆる形式の最新データを、社内チームだけでなく、顧客やパートナーとも共有して利用し、最大限の価値を引き出すデータ管理体制です。

これを実現するために企業は、自組織やデータプロバイダーをそれぞれのビジネスに最適なデータにつなぐための、一元化されたグローバルシステムを必要とします。効果的なソリューションは、構造化データ、半構造化データ、非構造化データ全てを包括し、パブリッククラウド上でそれら全てのデータを保存、処理、分析できる単一でシームレスなエクスペリエンスを提供する必要があります。

本eBookに掲載されているベストプラクティスは、すべてのデータの価値を今すぐ最大化するための有用なヒントとなります。非構造化データを単一のデータプラットフォームで保存、アクセス、管理、共有する奉納に関する詳細については、当社のウェビナー、「[7 Ways to Start Using Unstructured Data in Snowflake](#) (Snowflakeで非構造化データの利用を開始する7つの方法)」をご覧ください(非構造化データのサポートは現在プレビュー版です)。





Snowflakeについて

Snowflakeは、Snowflakeのデータクラウドを用い、あらゆる組織が自らのデータを活用できるようにします。お客様には、データクラウドを利用してサイロ化されたデータを統合し、データを発見してセキュアに共有し、多様な分析ワークロードを実行していただけます。データやユーザーがどこに存在するかに関係なく、Snowflakeは複数のクラウドと地域にまたがり単一のデータ体験を提供します。多くの業界から何千ものお客様（2021年10月31日時点で、2021年Fortune 500社のうちの223社を含む）が、Snowflakeデータクラウドを自社のビジネスの向上のために活用しています。詳しくは、[snowflake.com](https://www.snowflake.com)をご覧ください



© 2022 Snowflake Inc. All rights reserved. Snowflake、Snowflakeのロゴ、および本書に記載されているその他すべてのSnowflakeの製品、機能、サービス名は、米国およびその他の国におけるSnowflake Inc.の登録商標または商標です。本書で言及または使用されているその他すべてのブランド名またはロゴは、識別目的でのみ使用されており、各所有者の商標である可能性があります。Snowflakeが、必ずしもかかる商標所有者と関係を持ち、または出資や支援を受けているわけではありません。

参考

1 bit.ly/3if52aK

2 wikipedia.org/wiki/Semi-structured_data

3 bit.ly/2XQufkz

4 wikipedia.org/wiki/Unstructured_data

5 bit.ly/3ClqnkS

6 bit.ly/2WlxxvU

7 bit.ly/3if6V7A

8 gtnr.it/2XSRzOC

9 vz.to/3zLjyx8